AD_____

Award Number: DAMD17-96-1-6229

TITLE: Cancer Missed on Mammography

PRINCIPAL INVESTIGATOR: Robert M. Nishikawa, Ph.D.

CONTRACTING ORGANIZATION: University of Chicago
Chicago, Illinois 60637

REPORT DATE: October 2004

TYPE OF REPORT: Final

PREPARED FOR: U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

DISTRIBUTION STATEMENT: Approved for Public Release;
Distribution Unlimited

The views, opinions and/or findings contained in this report are those of the author(s) and should not be construed as an official Department of the Army position, policy or decision unless so designated by other documentation.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. AGENCY USE ONLY (Leave blank)	2. REPORT DATE	3. REPORT TYPE AND DATES COVERED
	October 2004	Final (23 Sep 1996 - 22 Sep 2004)

4. TITLE AND SUBTITLE
Cancer Missed on Mammography

5. FUNDING NUMBERS
DAMD17-96-1-6229

6. AUTHOR(S)
Robert M. Nishikawa, Ph.D.

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
University of Chicago
Chicago, Illinois 60637

E-Mail: r-nishikawa@uchicago.edu

8. PERFORMING ORGANIZATION REPORT NUMBER

9. SPONSORING / MONITORING AGENCY NAME(S) AND ADDRESS(ES)
U.S. Army Medical Research and Materiel Command
Fort Detrick, Maryland 21702-5012

10. SPONSORING / MONITORING AGENCY REPORT NUMBER

11. SUPPLEMENTARY NOTES

12a. DISTRIBUTION / AVAILABILITY STATEMENT
Approved for Public Release; Distribution Unlimited

12b. DISTRIBUTION CODE

13. ABSTRACT (Maximum 200 Words)

Observer error in reading screening mammograms has been identified as a significant factor in delayed diagnosis of breast cancer, with up to 30% of potentially detectable cancers being overlooked. Computer-aided detection (CADe) has been developed to aid radiologists in the detection task, and pre-clinical studies have shown that CADe applied to digitized mammography films can flag about 50% of radiologists' observational oversights. The purpose of this investigation is to test how many additional cancers are detected by radiologists using CADe, in an observer study using an enriched mixture of cancers. Based on an initial pilot study, we estimate that we need 360 cases containing 60 cancers with 10 radiologists. The CADe system used in this study has a sensitivity of 53% with 0.48 false positives per image for the cases used in this study. To date five radiologists have completed the study. All were experienced radiologists spending 100% of their clinical time on breast imaging. For these readers, we did not measure a statistically significant increase in performance when using CADe. Several less experienced radiologists are now enrolled in the study. We will be able to examine the effect of CADe as a function of reader experience.

14. SUBJECT TERMS
Mammography, computer-aided detection, missed cancers

15. NUMBER OF PAGES
46

16. PRICE CODE

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified	Unclassified	Unlimited

Table of Contents

Cover 1

SF 298 2

Table of Contents 3

Introduction 4

Body 4

Key Research Accomplishments 11

Reportable Outcomes 12

Conclusions 13

References 13

Appendices 15

4. INTRODUCTION

Double reading of mammograms has been shown to significantly increase the number of cancers detected [Beam 1996; Destounis 2004; Duijm 2004; Kopans 2000; Metz 1992; Thurfjell 1994; Warren 1995]. Computer-aided detection (CADe) has been proposed as an efficient method of implementing double reading [Schmidt 1994]. For CADe to be effective computers must find cancers that are missed by radiologists, and radiologists must react appropriately to the computer prompts. Others and we have found that CADe schemes can find over 50% of the observational misses made by radiologists reading mammograms [Birdwell 2001; Burhenne 2000; Nishikawa R M 2001; Schmidt 1996; te Brake 1998]. Our current study is designed to show that CADe can help detect cancers that they might otherwise be overlooked. That is, we will determine what fraction of cancers initially missed by a radiologist and will be detected by the radiologist when the cancers is flagged by the CADe scheme. We have collected a large database of cancers already missed by radiologists in routine clinical practice, and are testing observers without and with the aid of CADe in a controlled observer study.

5. BODY OF REPORT

5.1 Tasks

There are five tasks in the Statement of Work, which are listed below.

Task 1. Preparation of review forms and finalization of eligibility characteristics for cases to be entered into the missed lesion database.

Task 2. Accumulation of database cases and copying/digitizing 100 missed malignant cases and 300 normal cases, with categorization of features and characteristics of the malignant case. Verification of missed lesion cases.

Task 3. Computer runs producing hard copy of computer output for use in observer experiment and preparation of cases for observer experiment. Final design of details of observer performance study. This included a pilot study that was not in our original Statement of Work.

Task 4. An observer study comparing radiologists' performance in detecting breast cancer in screening mammograms with and without computer aid.

Task 5. Final analysis of data comparing CADe observer results with non-CADe results and observer variability, and preparation of report summarizing the results of the observer experiment and the clinical characteristics of the missed lesions.

5.1.1 Preparation of forms

A copy of the review is attached. The eligibility criteria are as follows:

1. Patients who have had screen-film mammograms read at the participating mammography facilities.

2. For cases of missed lesions, the mammogram had to be read clinically as normal in the area where a cancer subsequently developed, and the error had to be one of observation (failure to see the lesion) rather than interpretation (seeing the lesion and categorizing it as benign). In cases where the cancer is visible on multiple examinations prior to diagnosis, the two expert mammographers reviewing the cases will collaboratively select a single representative screening exam as the index missed case.

3. Case is a minimum of 1 year old (to avoid any interference with clinical care), unless bilateral mastectomy has been performed, or unless films clinically equivalent to those entered into the study from other years are available.

4. Case is not involved in any medical-legal action.

5. No copy films will be used that include significant marks made by a previous observer prior to the copying, and no originals with such permanent marks will be used.

6. Films from a prior exam will be collected and used in the study. If multiple prior exams are available, an exam that is either 1 or 2 years old will be used.

5.1.2 Development of database of missed lesions

The full database consists of 360 normal cases and 75 cases with a missed cancer and 25 cases with a clinically detected cancer (for training purposes). Each case consisted of a current exam and a previous exam. All cases were screening mammograms consisting of two views (craniocaudal and mediolateral-oblique) of each breast. Tables 1-4 summarize the characteristics of the cancers in the database.

Table 1. Distribution of breast density in our database

BREAST DENSITY	FREQUENCY OF OCCURRENCE
Normal	0.30
Fatty	0.21
Dense	0.37
Focal	0.09

Table 2. Distribution of subtlety on a 5-point scale, where 1 is extremely subtle.

SUBTLETY RATING	FREQUENCY OF OCCURRENCE
1	0.16
2	0.39
3	0.37
4	0.05
5	0

Table 3. Distribution by lesion type*

TYPE OF LESION	FREQUENCY OF OCCURRENCE
Asymmetric Density	0.29
Architectural Distortion	0.24
Developing Density	0.07
Mass	0.46
Calcifications	0.10

*numbers sum to greater than 1, because

Table 4. Distribution of possible reasons for cancers being missed.

POSSIBLE REASON	FREQUENCY OF OCCURRENCE
Seen on only 1 view	0.48
Obscured by overlying tissue	0.40
Looks like normal tissue	0.36
"Busy" breast	0.29
Film technique	0.26
Distracting lesions	0.24
Subtle lesion	0.14
Marginal lesion	0.10
Developing density	0.10
Benign appearing lesion	0.07
Lack of prior films	0.07
Too small to prompt workup	0.05
Lucent lines	0.05
Stable lesion	0.02

*numbers sum to greater than 1, because up to three reasons were given per case.

5.1.3 Computer analysis of case

To determine the number of cases and readers we need in our formal observer study [Nishikawa R. M. 2001], we conducted a pilot study (a reprint is attached). In a prospective evaluation of computer detection schemes developed in our laboratory, we have analyzed over 12,000 clinical mammographic screening exams. Retrospective review of the negative screening mammograms for all cancer cases found an indication of the cancer in 23 of these negative cases. The computer found 54% of these in our prospective testing. We added to these cases normal exams to create a dataset of 75 cases. Four radiologists experienced in mammography read the

cases and gave their BI-RADS assessment and their confidence that the patient should be called back for diagnostic mammography. They did so once reading the films only and a second time reading with the computer aid. Three radiologists had no change in area under the ROC curve (mean A_z of 0.73) and one improved from 0.73 to 0.78, but this difference failed to reach statistical significance (p=0.23).

From this pilot study, we determined that the correlation in A_z values between aid and unaided reading conditions was between 0.82 and 0.99, with an average of 0.93 (see Table 5). Using a conservative correlation value of 0.82, we estimate for a single reader that we would need approximately 400 cases that included 80 cancers to have 80% power to measure a difference in A_z of 0.06. If we assume a correlation value of 0.93, then 200 cases with 40 cancers would give 80% power.

We initially planned to use 400 cases and 80 cancers. However, Astley *et al.* reported at a conference in the summer of 2004, that extensive training of radiologists in using CADe is necessary in order for the radiologists to use the computer aid effectively. Therefore, we originally had planned a small training set (n=20), principally to allow the radiologists to become use to using our computer interface used in the study to display the computer results and to collect the radiologists' interpretations. We now require the radiologists to review 40 training cases for session 1 and 20 training cases in each of the other three sessions for a total of 100 training cases. Note, that this is still less than the approximately 400 cases that Astley found necessary for proper training. However, we do not have the resources to use 400 training cases. Further, the Astley result is unverified. Therefore, we used 300 cases with 66 cases containing 69 cancers in our study. There were 234 cases that did not contain a cancer.

Table 5. Summary of reader performance from pilot observer study.

Reader	A_z Unaided	A_z With Aid	Correlation between A_z aid and no aid
A	0.686	0.685	0.967
B	0.725	0.775	0.817
C	0.805	0.793	0.943
D	0.710	0.688	0.988
mean	0.731	0.735	0.929

We determined from the pilot study that CADe schemes with fewer false positives is needed for our study. Our current detection schemes (one for masses and one for calcifications) have about 3 false positives per image. Commercial systems average under 1 per image. Note that because of the success of commercial software, we have not developed our detection since approximately 1998. We have on loan, an R2 Technology, Inc, ImageChecker 1000 that has a sensitivity of approximately 85% with 0.5 false positives per image [Roehrig 1998]. We will used this system in our study. All the cases have been analyzed by the R2 System. For the 69 cancers used in the reader study, the computer's sensitivity was approximately 53%. The false positive rate was 0.48 per image.

5.1.4 Observer study

The formal observer study is underway, but is not finished. Ten radiologists will perform the study. Each reader will be asked to answer 2 questions for each case:

> If you were reading this case clinically and this is all the information that is available: (i) Give your BI-RADS assessment of this case; and (ii) what is your level of confidence that the patient should be called back for further work-up or a biopsy? Answer the second question using the following confidence scale:
>
> 1.0 No evidence for recalling the patient.
>
> 1.5
>
> 2.0 Some, but insufficient evidence for recalling the patient
>
> 2.5
>
> 3.0 Equivocal. [If you read this case on 10 different days, half the time you would recall.]
>
> 3.5
>
> 4.0 Sufficient evidence for recalling the patient.
>
> 4.5
>
> 5.0 Overwhelming evidence for recalling the patient.

If the radiologist gives a BI-RADS assessment is not 1, then the radiologist will be required to specify the location of the lesion and type of lesion using the computer interface.

The BI-RADS rating and lesion type and location will be used to generate sensitivity and call back rates, while the confidence scale will be used to do ROC analysis.

The 300 cases were divided into 4 groups, so that each reader needed to complete 4 reading sessions. Reading session 1 contained 40 training cases and 60 study cases. The remaining 3 sessions each had 20 training cases and 80 study cases. The films were hung on a motorized viewer designed especially for mammography. The room lights were turned off. The radiologists had available to them magnifying glasses and a bright light. No time limit was imposed. A sequential reading process is used. Specifically, each reader reviews the film without any computer aid and then answers the two questions listed above. After answering the questions, the reader is shown the CADe detections via a computer interface that shows minified

versions of the films being read. The reader then answers the two questions again incorporating any CADe findings.

5.1.5 Data Analysis

For the five readers who have complete the study, we performed ROC analysis using e the Dorfman, Berbaum, Metz method (also known as MRMC method – multiple readers, multiple cases) for testing the statistical significance of differences in the area under the ROC curve [Dorfman 1992]. We found no statistically significant difference between the aided and unaided reading conditions – the 95% confidence interval for the difference in Az is (-0. 0595,0.0213). As seen in Table 6, 3 readers increased their Az while 2 readers had a decrease. All readers were full breast imagers with over 10 years of experience. We are now testing less experience readers.

Table 6. Performance of the first 5 radiologists.

Reader	Az Unaided	Az With Aid	95% Confidence Intervals for difference in Az
A	0.7786	0.7800	(-0.0281,0.0252)
B	0.7371	0.8039	(-0.1592,0.0256)
C	0.7267	0.7689	(-0.1098,0.0255)
D	0.7670	0.7555	(-0.0023,0.0252)
E	0.8164	0.8131	(-0.0149,0.0216)
mean	0. 7652	0. 7843	(-0. 0595,0.0213)

5.2. Recommendations in relation to the Statement of Work

We made two major changes to the original statement of work:

1. We originally planned to use CADe schemes developed in our laboratory. We instead used a commercial system because it had better performance than our own schemes. We concluded from our pilot study that the false positive rate of our CADe schemes were too high, because it appeared that radiologists were not recognizing missed cancers flagged by the computer.

2. We changed the number of cases and readers based on the pilot study and we increased the number of training cases based on unpublished reports on the effect of training on CADe use.

5.3. Discussion

There have been three clinical studies involving CADe published [Destounis 2004; Freer 2001; Gur 2004]. Two have found that CADe helps detect more cancers (although the increase was not statistically significant) [Destounis 2004; Freer 2001]and one, the largest study, did not find that CADe increase the cancer detection rate [Gur 2004]. It is important to conduct controlled studies to try to understand the possible reasons for this possible discrepancy. In clinical studies, it is difficult to draw conclusions because only a single radiologist reads each case and there exists large variation between readers [Schmidt 1998]. In an observer study, multiple readers view the same case allowing for better statistical power in data analysis. For example, although all these cases were missed clinically, many readers detected some of the misses. Further there are cases in which the computer identified the cancer, but some readers still not detect the cancer when using the computer aid. If there are a majority of readers detected the cancer, but some readers ignore the computer prompt then different conclusions can be drawn then if none or a small number of the readers detected the cancer.

We will perform an in-depth analysis of the observer study when all readers have finished. This will include:

- performance of individual radiologists with and without aid
- number of additional cancers detected and the increase in callback rate when computer aid is used
- radiologists' performance with and without CADe broken down by lesion type
- radiologists' performance with and without CADe broken down by the radiologists' experience
- radiologists' performance on individual cases broken down computer's accuracy (i.e., whether the computer detected the cancer and the number of false positives)
- radiologists' performance with and without CADe broken down by the reason why the cancer was missed clinically
- computer performance broken down by the reason why the cancer was missed clinically

6. KEY RESEARCH ACCOMPLISHMENTS

- Pilot observer study performed

- Detailed planning of observer study. This is the first observer study of screening mammography to include previous exams. This makes the study more clinically realistic and thus the results more applicable to clinical mammography.

- Formal observer study half completed. We expected to have the full study completed by the March 2005.

7. REPORTABLE OUTCOMES

Based on the support from this grant we have published 4 conference proceeding papers, have 1 paper in preparation for a peer-reviewed journal, have given 5 talks at international meetings, and three scientific posters, two of which won a Cum Laude award.

Conference Proceedings

1. Schmidt RA, Newstead GM, Linver MN, Eklund GW, Metz CE, Winkler MA, Nishikawa RM: Mammographic screening: Sensitivity of general radiologists. In: Karssemeijer N, Thijssen M, Hendriks J and van Erning L (eds.), Digital Mammography. (Amsterdam: Kuwester) 1998, pp. 383-388.
2. Nishikawa RM, Giger ML, Wolverton DE, Schmidt RA, Comstock CE, Papaioannou J, Collins SA, Doi K: Prospective testing of a clinical mammography workstation for CAD: Analysis of the first 10,000 cases. In: Karssemeijer N, Thijssen M, Hendriks J and van Erning L (eds.), Digital Mammography. (Amsterdam: Kuwester) 1998, pp. 401-406.
3. Nishikawa RM, Giger ML, Schmidt RA, Vyborny CJ, Bick U, Doi K: Prospective computer analysis of cancers missed on screening clinical. In: Digital Mammography 2000, Yaffe MJ, (ed). (Medical Physics Publishing, Madison WI) 2000, 493-498.
4. Nishikawa RM, Giger ML, Schmidt RA, Papaioannou J: Can computer-aided diagnosis (CAD) help radiologists find mammographically missed screening cancers? Proc. SPIE 4324:56-63, 2001.

Presentations

1. Nishikawa RM, Giger ML, Wolverton DE, Schmidt RA, Comstock CE, Papaioannou J, Collins SA, Doi K: Prospective testing of a clinical mammography workstation for CAD: Analysis of the first 10,000 cases. Presented at the Fourth International Workshop on Digital Mammography, June 1998, Nijmegen, The Netherlands.
2. Schmidt RA, Newstead GM, Linver MN, Eklund GW, Metz CE, Winkler MA, Nishikawa RM: Mammographic screening: Sensitivity of general radiologists. Presented at the Fourth International Workshop on Digital Mammography, June 1998, Nijmegen, The Netherlands.
3. Nishikawa RM, Giger ML, Schmidt RA, Wolverton DE, Collins SA, Doi K, *et al.* Computer-aided diagnosis in screening mammography: Detection of missed cancers. Presented at 84th Scientific Assembly of the Radiological Society of North America, November 1998, Chicago, IL.
4. Nishikawa RM, Giger ML, Schmidt RA, Vyborny CJ, Bick U, Doi K: Prospective computer analysis of cancers missed on screening clinical. In: Digital Mammography 2000, Yaffe MJ, (ed). (Medical Physics Publishing, Madison WI) 2000, 493-498.

5. Nishikawa RM, Giger ML, Schmidt RA, Papaioannou J: Can computer-aided diagnosis (CAD) help radiologists find mammographically missed screening cancers? Presented at SPIE Medical Imaging 2001, February 2001, San Diego CA.

Scientific Poster

1. Jiang Y, Nishikawa RM, Giger ML, Huo Z, Schmidt RA, Wolverton DE, *et al.*: Computer aided diagnosis of breast lesions: An interactive demonstration. 84th Scientific Assembly and Annual Meeting of the Radiological Society of North America, November 1998, Chicago, IL. (Awarded Cum Laude).
2. Giger ML, Nishikawa RM, Huo Z, Jiang Y, Venta LA, Doi K, Vyborny CJ: Computer-aided diagnosis (CAD) in breast imaging. 85th Scientific Assembly and Annual Meeting of the Radiological Society of North America, November 1999, Chicago, IL.
3. Nishikawa RM, Giger ML, Jiang Y, Huo Z, Vyborny CJ, Jokich P: Implementation of computer-aided diagnosis into the clinical mammography work flow. 86th Scientific Assembly and Annual Meeting of the Radiological Society of North America, November 2000, Chicago, IL (Awarded Cum Laude).

8. CONCLUSIONS

The goal of this project was to show that a computer can alert radiologists to missed cancers. The observer study to prove this is half completed so definitive conclusions cannot be made at this time. We have found, in preliminary analysis, a slight increase in performance for experience radiologists and we are now testing less experienced radiologists.

9. REFERENCES

Beam CA, Sullivan DC, and Layde PM. 1996. Effect of human variability on independent double reading in screening mammography. *Academic Radiology* **3**: 891-897.

Birdwell RL, Ikeda DM, O'Shaughnessy KF, and Sickles EA. 2001. Mammographic characteristics of 115 missed cancers later detected with screening mammography and the potential utility of computer-aided detection. *Radiology* **219**: 192-202.

Burhenne LJW, Wood SA, D'Orsi CJ, Feig SA, Kopans DB, O'Shaughnessy KF, Sickles EA, Tabar L, Vyborny CJ, and Castellino RA. 2000. Potential contribution of computer-aided detection to the sensitivity of screening mammography. *Radiology* **215**: 554-562.

Destounis SV, DiNitto P, Logan-Young W, Bonaccio E, Zuley ML, and Willison KM. 2004. Can computer-aided detection with double reading of screening mammograms help decrease the false-negative rate? Initial experience. *Radiology* **232**: 578-584.

Dorfman DD, Berbaum KS, and Metz CE. 1992. Receiver operating characteristic rating analysis. Generalization to the population of readers and patients with the jackknife method. *Investigative Radiology* **27**: 723-31.

Duijm LEM, Groenewoud JH, Hendriks JHCL, and Koning HJd. 2004. Independent double reading of screening mammograms in the Netherlands: Effect of arbitration following reader disagreements. *Radiology* **251**: 564-570.

Freer TW, and Ulissey MJ. 2001. Screening mammography with computer-aided detection: prospective study of 12,860 patients in a community breast center. *Radiology* **220**: 781-786.

Gur D, Sumkin JH, Rockette HE, Ganott M, Hakim C, Hardesty L, Poller WR, Shah R, and Wallace L. 2004. Changes in breast cancer detection and mammography recall rates after the introduction of a computer-Aided detection system. *Journal of the National Cancer Institute* **96**: 185-190.

Kopans DB. 2000. Double reading. *Radiologic Clinics of North America* **38**: 719-24.

Metz CE, and Shen JH. 1992. Gains in accuracy from replicated readings of diagnostic images: prediction and assessment in terms of ROC analysis. *Medical Decision Making* **12**: 60-75.

Nishikawa RM, Giger ML, Corbridge C, Schmidt RA, and Papaioannou J. 2001. Can computer-aided diagnosis help radiologists find mammographically missed screening cancers? *Proceeding SPIE* **4324**: 56-63.

Nishikawa RM, Giger ML, Schmidt RA, Vyborny CJ, Bick U, and Doi K. 2001. Prospective computer analysis of cancers missed on screening clinical. In: *Digital Mammography 2000.* Yaffe MJ, Eds. (Medical Physics Publishing, Madison WI),

Roehrig J, Doi T, Hasegawa A, Hunt B, Marshall J, Romsdahl H, Schneider A, Sharbaugh R, and Zhang W. 1998. Clinical results with R2 Imagechecker system. In: *Digital Mammography.* Karssemeijer N, Thijssen M, Hendriks J, and van Erning L, Eds. (Kluwer Academic Publishers, Boston), pp. 395-400.

Schmidt RA, Newstead GM, Linver MN, Eklund GW, Metz CE, Winkler MA, and Nishikawa RM. 1998. Mammographic screening: Sensitivity of general radiologists. In: *Digital Mammography - Nijmegen, 1998.* Karssemeijer N, Thijssen M, Hendriks J, and Erning Lv, Eds. (Kluwer Academic Publishers, Dordrecht, The Netherlands), pp. 383-388.

Schmidt RA, and Nishikawa RM. 1994. Digital screening mammography. *Principles & Practice of Oncology* **8**: 1-16.

Schmidt RA, Nishikawa RM, Osnis R, Schreibman KL, Giger ML, and Doi K. 1996. Computerized detection of lesions missed by mammography. In: *Digital Mammography '96.* Doi K, Giger ML, Nishikawa RM, and Schmidt RA, Eds. (Elsevier Science, Amsterdam), pp. 105-110.

te Brake GM, Karssemeijer N, and Hendriks JHCL. 1998. Automated detection of breast carcinomas not detected in a screening program. *Radiology* **207**: 465-471.

Thurfjell EL, Lernevall KA, and Taube AAS. 1994. Benefit of independent double reading in a population-based mammography screening program. *Radiology* **191**: 241-244.

Warren RML, and Duffy SW. 1995. Comparison of single reading with double reading of mammograms, and change in effectiveness with experience. *British Journal of Radiology* **68**: 958-962.

10. APPENDICES

Four reprints are attached:

1. Schmidt RA, Newstead GM, Linver MN, Eklund GW, Metz CE, Winkler MA, Nishikawa RM: Mammographic screening: Sensitivity of general radiologists. In: Karssemeijer N, Thijssen M, Hendriks J and van Erning L (eds.), Digital Mammography. (Amsterdam: Kuwester) 1998, pp. 383-388.
2. Nishikawa RM, Giger ML, Wolverton DE, Schmidt RA, Comstock CE, Papaioannou J, Collins SA, Doi K: Prospective testing of a clinical mammography workstation for CAD: Analysis of the first 10,000 cases. In: Karssemeijer N, Thijssen M, Hendriks J and van Erning L (eds.), Digital Mammography. (Amsterdam: Kuwester) 1998, pp. 401-406.
3. Nishikawa RM, Giger ML, Schmidt RA, Vyborny CJ, Bick U, Doi K: Prospective computer analysis of cancers missed on screening clinical. In: Digital Mammography 2000, Yaffe MJ, (ed). (Medical Physics Publishing, Madison WI) 2000, 493-498.
4. Nishikawa RM, Giger ML, Schmidt RA, Papaioannou J: Can computer-aided diagnosis (CAD) help radiologists find mammographically missed screening cancers? Proc. SPIE 4324:56-63, 2001.

MAMMOGRAPHIC SCREENING: SENSITIVITY OF GENERAL RADIOLOGISTS

ROBERT A. SCHMIDT[1], GILLIAN M. NEWSTEAD[2], MICHAEL N. LINVER[3], G.W. EKLUND[4], CHARLES E. METZ[1], MICHAEL A. WINKLER[1], AND ROBERT M. NISHIKAWA[1]

[1] *Kurt Rossmann Laboratories for Radiologic Image Research, Department of Radiology, The University of Chicago, Chicago, IL, USA*
[2] *Dept. of Radiology, New York University, NY, NY, USA*
[3] *Dept. of Radiology, University of New Mexico at Albuquerque, and X-Ray Associates of New Mexico, Albuquerque, NM, USA*
[4] *University of Illinois College of Medicine at Peoria, Peoria, IL, USA*

1. Introduction

High quality mammography can detect early, curable breast cancer and decrease mortality. Much research effort is being expended to improve mammography (digital mammography, computer-aided diagnosis [CAD]), and develop alternative modalities (ultrasound, MRI, radionuclide imaging). However, the human observer is at this point potentially the weakest link in the diagnostic imaging chain, and the range of performance in routine practice is unknown. We have conducted a large observer study using a standardized test base to further investigate this issue.

2. Materials and methods

We did our study at five meeting locations in the US in 1997 and 1998, using films selected from three clinical mammography practices. High quality copy films of 100 cases were presented to a total of over 250 observers who were attending these conferences, and 4 selected experts. Films were displayed (without prior studies or clinical history) on motorized viewboxes designed for mammography, and observers given about 2 1/2 hours to complete the exercise, in supervised workshop settings that typically had 1 to 3 radiologists per viewbox. The case mix was 45 cases containing 50 cancers diagnosed in routine practice, and 55 normal/benign cases. Data were collected regarding the level of experience of observers and the number of mammograms they read. We have graded the first 100 observers for this report.

The distribution of breast lesions in the test set mirrored that in clinical practice, with an emphasis on masses, distortions and asymmetries, rather than calcifications. Microcalcifications probably account for 40 to 50% of tissue sampling breast interventional procedures in the US, but their average positive biopsy yield is lower than that of masses, particularly after the introduction of less invasive percutaneous needle sampling techniques. The perceptual problems in screening associated with detection of significant soft tissue abnormalities is considered harder than the detection of microcalcifications by the authors of this paper, and invasive cancers are more life threatening; hence the emphasis on this type of potentially missed lesion. The distribution of morphologies on mammogram of the breast cancers was: spiculated mass (Mass-S) - 42%, circumscribed mass (Mass-C) - 6%, architectural distortion (ARD) - 14%, asymmetric density (ASD) - 6%, mass + calcifications - 6%, mass + ARD - 6%, ASD or ARD + calcifications - 10%, microcalcifications only (Ca^{++}) - 10%. Figure 1 shows the relative number of lesions of different types, graded by our assessment of their mammographic suspicion (BIRADS-type rating, with 5 being the highest suspicion).

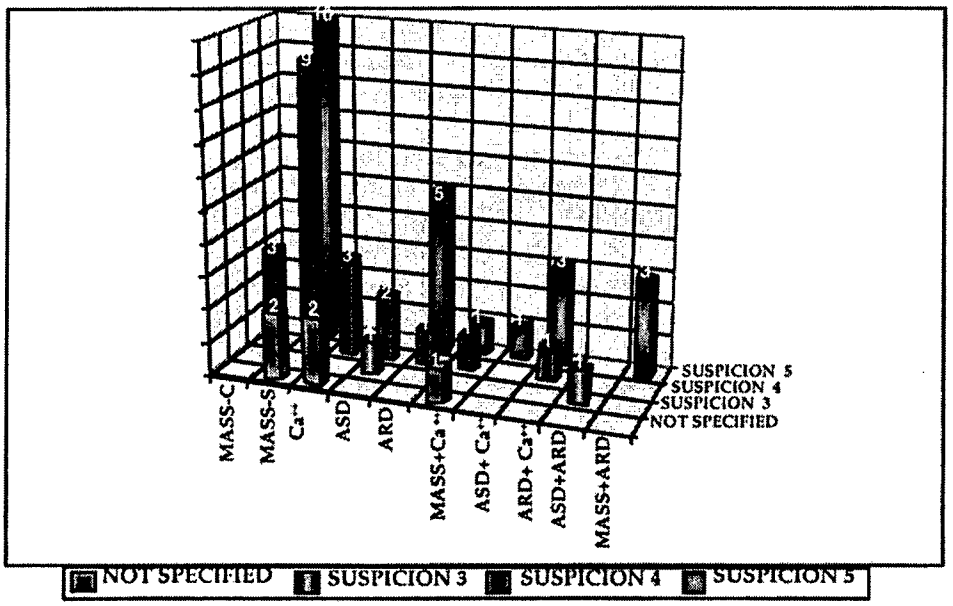

Figure 1. Distribution of lesion types by mammographic suspicion. The terms are defined in the text.

84% of the cancers were invasive, and 16% in situ. The distribution of pathologic types of the breast cancers was: infiltrating ductal carcinoma (IDC) - 48%, IDC + ductal carcinoma in situ (DCIS) - 12%, infiltrating lobular carcinoma (ILC) - 12%, IDC/ILC -

2%, tubular IDC - 6%, medullary IDC 2%, papillary IDC 2%, pure DCIS - 16%. The cancers chosen were representative of average difficulty cases encountered in routine screening practice. They were not the most subtle or tricky cases, and none were cancers which had been "missed," although a typical fraction had films prior to the index (detected) case where the cancer could have been diagnosed. These prior films were not shown to the observers. At the first session only, previous mammograms were hung above the test cases for comparison, but this slowed down the observers, and made it difficult to maintain the pace needed to complete the exercise in a reasonable time.

Mammograms were from examinations taken since 1986, with the majority comparable in technical quality to the range of examinations seen in current clinical practice. The same films were shown to all participants. All except one case (a unilateral examination) had two views of each breast, hung on RADX Mammoscope viewers brought to the meeting, with MLO and CC views hung with right and left views back to back. Light restricting shutters were used, room lights dimmed and magnifying lenses made available, to simulate normal clinical practice Approximately 12 cases were hung on each of 8 automated viewers. A two part NCR carbonless form was devised for scoring (figure 2), so that participants could retain one copy while going over the answers to the cases with an expert at the viewboxes, at the end of the session. This also ensured that answers were not altered at the time of review. Observers were asked to mark whether the case was normal (corresponding to BIRADS codes 1 and 2) or abnormal. If an abnormality was detected, they were asked to mark the lesion type, location on two views, if possible, and their level of suspicion on a five point scale. In subsequent work, we have used a 10 point suspicion scale, to generate ROC curves. Readers were told there were more normal than abnormal cases, and were given about 1 minute per case, with the structured exercise lasting 2 1/2 hours. An additional 1 1/2 hours were devoted to going over the individual cases with participants in small groups.

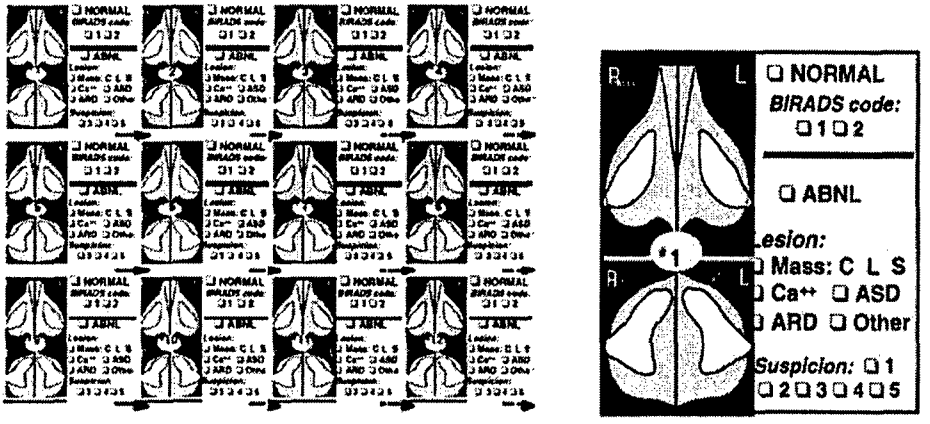

Figure 2. Sample scoring sheet, and enlargement to show detail

If there was more than one lesion, the observers were asked to mark these "#1," "#2," etc. The case set did contain an enriched mixture of bilateral cancers: 5 cases out of 45 patients with cancer, or 11%. Synchronous bilateral cancer is usually only seen in about 3% of screening detected cancers, but we knew from previous work that such cases are disproportionately represented in series of cases that radiologists miss.

Answers were graded correct if the case was marked abnormal and the correct location of the cancer was marked on at least one view, judged as the mark within a distance approximately 1/3 of the distance between the nipple and the chest wall from the true location. If a positive case was left blank, a false negative was scored. If a negative case was left blank, a false positive was scored. Only participants who answered more than 90% of the 105 "cases" (55 normals, and 45 abnormals containing 50 cancers) were considered to have completed the exercise. This eliminated 25% of the 100 readers, and 1 of the 4 experts, who were thereby dubbed "non-compliers." Results are given in Figure 3 for those who completed the test, and in Figure 4 separately for the 22 non-complying radiologists (and 3 non-radiologists). For the latter figure, however, only the cases for which an answer was given were graded. On average, experts, complying radiologists and non-compliers answered 99.7%, 97.3% and 79.1% of the cases, respectively.

3. Results

Based on analysis of the first 100 observers and four experts, there is a considerable range of accuracy in reading screening mammography among general radiologists in the US, and expert mammographers are generally better at the screening task (Figure 3):

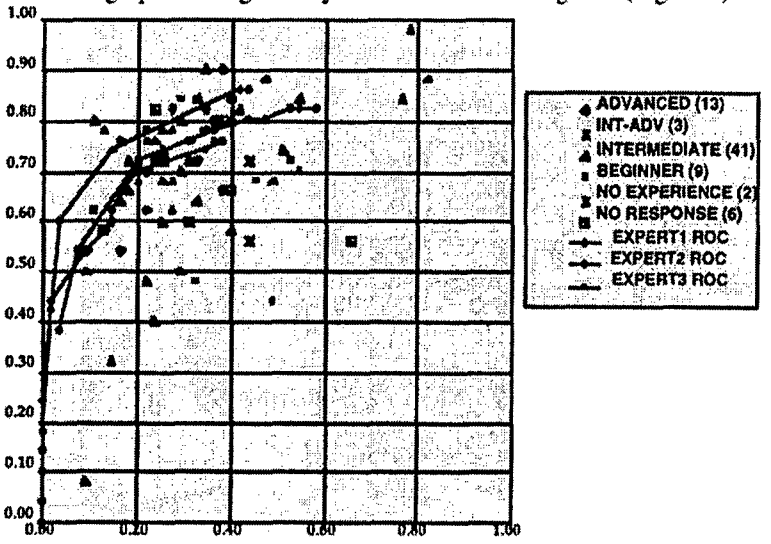

Figure 3. Scatter plot of true positive fraction [TPF, ordinate] versus false positive fraction [FPF, abscissa] for 75 general radiologists, and ROC curves for 3 experts

For the observers who gave answers to more than 90% of the cases, the average sensitivity was 70%, with range of 8% to 98% (SD 14%; spec = 68%, with range of 18 to 91%) for correct cancer detection and localization for 75 general radiologists, and 81% for 3 experts, with smaller range (76% to 86%; specificity = 54%). Standard error of the mean was 3% in both cases. There was only a relatively weak correlation with general observers' self-assessment of their level of expertise (Figure 5). The sensitivity of those who did not complete 90% of the cases was only 42%, with specificity of 72%.

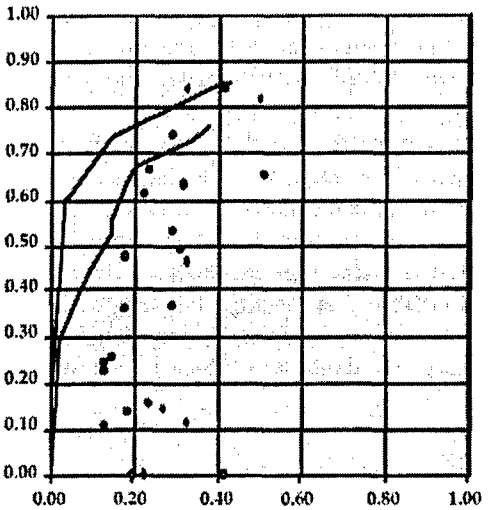

Figure 4. Scatter plot of TPF versus FPF for the 25 "non-complying" radiologists, and the ROC curves of the best and worst of 3 experts

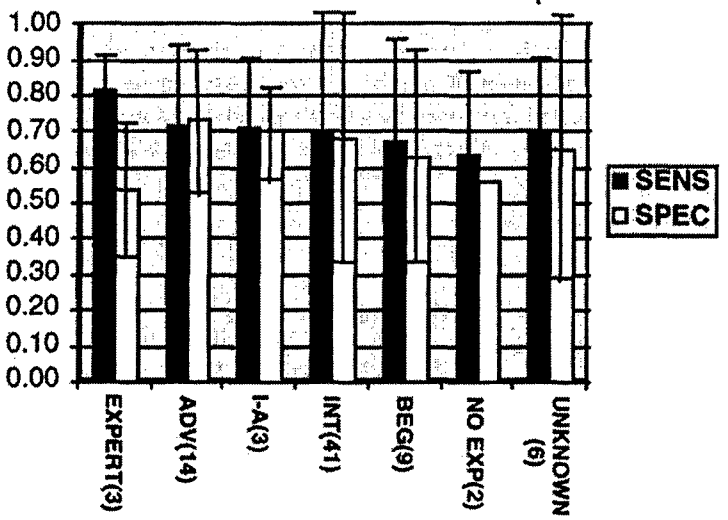

Figure 4. Sensitivity and specificity, based on radiologists' self assessment. Experts were designated by the authors. Error bars are two standard deviations.

4. Discussion

This exercise was modeled after the second level Teaching Course in Mammography of László Tabár. We individually scored the participants (generating 30,000 data points), and they were given the option of getting specific quantitative feedback on their performance; more than half took this option. A certificate for reading 100 cases under supervision can be given towards MQSA requirements in the US.

While the results may not be unexpected, the range of performance in detecting breast cancers on screening mammography by general radiologists is quite large, and radiologists who are experts and dedicated to mammography perform substantially better than the average radiologists, detecting about 16% more cancers in this study. This increased sensitivity comes at the price of decreased specificity, however. This increase in detection rate is comparable to improvements that can be confidently expected from improvements in the mammogram images themselves, or by developing alternative modalities. The introduction of computer-aided diagnosis (CAD) techniques into clinical practice would be expected to decrease the gap between the average reader and the expert reader, and decrease the variability of readings, but this will require further large scale studies. There are also obvious implications for improving the training of radiologists, and establishing competency standards, which have not yet been implemented in the US.

5. Conclusions

General radiologists read mammograms with higher specificity and lower sensitivity than experts. There is room for improvement in breast cancer detection: experts are at least 16% more sensitive than general radiologists, and the variability of general radiologists is very high. There is a need for improved training and feedback for radiologists, with indication of a need for minimum competency testing. Benefits similar to those expected from imaging technology advances are likely possible, and one way that performance may be improved through technical advances is by use of CAD.

Acknowledgments

This work was supported in part by a grant from the US Army (DAMD17-96-6229). The contents of this paper are solely the responsibility of the authors and do not necessarily represent the official views of any of the supporting organizations. The authors gratefully acknowledge the assistance of David Joyner of RADX corporation, for providing us with the mammography viewstations used at each of the meeting locations, and of Brian Goode of Faxitron, in providing magnifying lenses for the participants for the last several sessions.

Reprint from "Digital Mammography. Nijmegen, 1998" N. Karssemeijer *et al.* editors. Kluwer Academic Publishers, Dordrecht, The Netherlands, 1998.

PROSPECTIVE TESTING OF A CLINICAL MAMMOGRAPHY WORKSTATION FOR CAD: ANALYSIS OF THE FIRST 10,000 CASES

ROBERT M. NISHIKAWA, MARYELLEN L. GIGER, DULCY E. WOLVERTON, ROBERT A. SCHMIDT, CHRISTOPHER E. COMSTOCK, JOHN PAPAIOANNOU, STEPHEN A. COLLINS, AND KUNIO DOI

Kurt Rossmann Laboratories for Radiologic Image Research, Department of Radiology, The University of Chicago
United States of America.

1. Introduction

For over ten years, we have been developing automated computerized schemes to assist radiologists in detecting breast cancer from mammograms. These detection schemes have been implemented on an "intelligent" mammography workstation that has been used prospectively on screening mammograms for over three years. The purpose of this study was to analyze the performance of the workstation in comparison to radiologists' clinical interpretations of the same screening mammograms.

2. Materials and methods

The clinical workstation consists of a Konica LD4500 film digitizer, an IBM RISC/6000 Powerstation 590 workstation, a Seikosha VP4500 thermal printer for hardcopy recording of the computer results, a 1600x1200 touchscreen monitor to enable radiologists to review the computer results during clinical use, and two automated detection schemes -- one for masses and the other for clustered microcalcifications. The touchscreen monitor was not used in this study. The automated detection schemes have been described previously and flowcharts are shown in Figure 1. Details of the detection schemes can be found in references [1-3] for the mass scheme and references [4-10] for the clustered microcalcifications scheme.

Since November 8, 1994, all screening mammograms taken at the University of Chicago Hospitals have been analyzed on the workstation, except during downtimes. Downtime has been minimal, less than 20 days in total, which includes a 3-week period when the mammography section moved to a new outpatient center. During that move, networking problems in the new facility contributed to computer system difficulties.

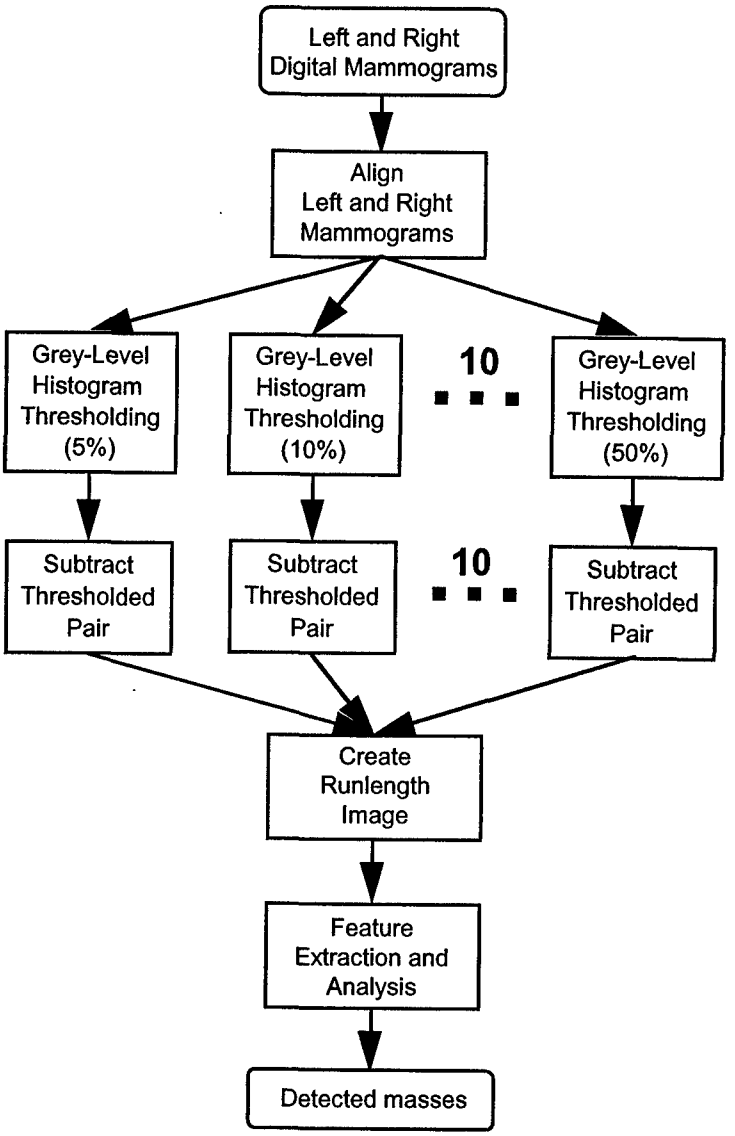

Figure 1. Flowchart for our automated scheme for the detection of masses on digital mammograms. Details about the technique can be found in references [1-3].

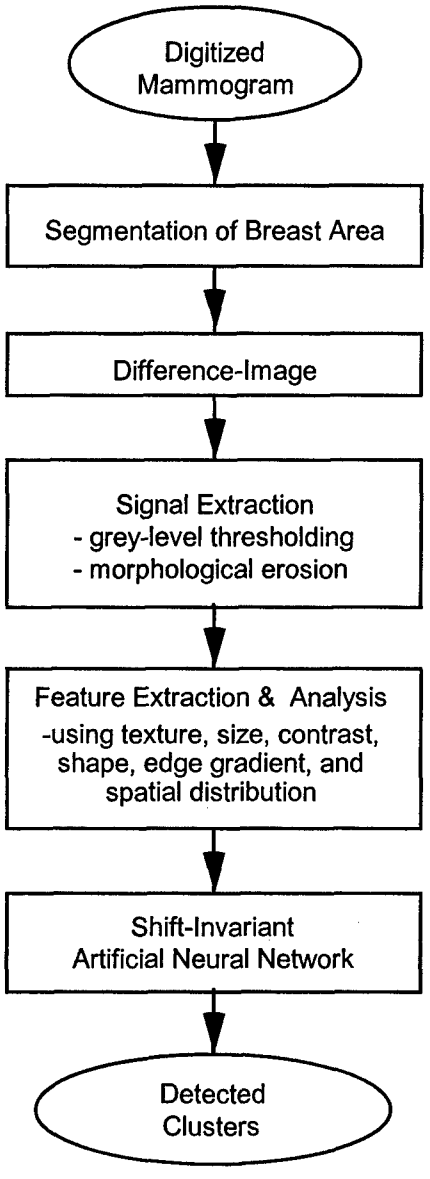

Figure 2. Flowchart for our automated scheme for the detection of clustered microcalcifications on digital mammograms. Details of the technique can be found in references [4-10].

For each case, four-view screening mammograms were digitized to 100-micron pixel size and 10-bit grey scale resolution. These were subjected to analyses by our two automated detection schemes. The performance of the computer was calculated in terms of sensitivity for detection of cancer and false-positive rate (detection of non-cancers per image).

For cases in which a cancer was detected, we also retrospectively reviewed any previous mammograms that were in our study cohort. Two radiologists independently reviewed the cases and stated whether the cancer was visible in a previous exam and whether, knowing that the lesion was present, that they would call the patient back for a diagnostic exam based on the findings in the previous exam. In this way, the number of cancers detected by the computer that were initially missed by the radiologists was determined.

3. Results

As of May, 1, 1998, over 14,000 cases have been analyzed. With follow-up on the first 10,000 cases, 61 patients have been diagnosed with breast cancer. In 12 of these cases, the screening mammogram(s) were negative even in retrospect. For the mammographically visible cases (n=49), the sensitivity of the two schemes was 68% (34/49). Clinically, 96% of the cancers were detected (47/49). More important than the absolute sensitivity of the workstation is its ability to detect breast cancers that may be missed by a radiologist. In 30 of the 61 cancers, the patient had a screening exam that was read as negative and was included in our study. That is, a screening mammogram that was read as normal, which preceded the cancer being diagnosed. In 14 of these cases, no lesion could be seen in retrospect, i.e., mammographically negative. In 9 of 16 cases, the computer was able to identify the region on the negative-read (cancer visible in retrospect) screening mammogram that corresponded to where the cancer was subsequently detected. Overall, the computer was able to identify the cancer approximately one year before it was diagnosed in approximately 15% (9/61) of all cancer cases and in 56% (9/16) of cases were the cancer was visible in retrospect on a negative-read screening mammogram. The false-positive rate was approximately 1.3 false clusters per image and 2.1 false masses per image.

4. Discussions

For CAD to be effective it must alert radiologists to cancers that the radiologist initially did not catch. The first step in accomplishing this is to detected cancers missed by the radiologist. The second step is to have the radiologist recognize that the computer detection is indeed a cancer. In this study we have examined the first step. We have found that, in a prospective study, the computer was able to detect approximately 50% of cancers that were initially overlooked but visible in retrospect. This is consistent with

our previous retrospective study [11] , and a recently published study by Karssemeijer *et al.* [12] .

Being able to detect cancers that radiologists can miss is only the one part of a successful implementation of CAD. It is also necessary that the radiologist act appropriately to the computer prompt, by calling the women back for further examination when a cancer is present and ignoring the prompt if a benign lesion has been detected. We are currently conducting a clinical evaluation of CAD to determine if radiologists can successfully use the assistance of our workstation.

5. Conclusions

In a prospective study, the automated detection schemes were able to identify approximately 50% of overlooked cancers and over 10% of all cancers approximately one year before diagnosis. With the large database being created, we can better optimize the performance of the system.

References

[1] Yin F-F, Giger ML, Doi K, Metz CE, Vyborny CJ, Schmidt RA (1991) Computerized detection of masses in digital mammograms: Analysis of bilateral subtraction images. Medical Physics 18, pp 955-963.

[2] Yin F-F, Giger ML, Doi K, Vyborny CJ, Schmidt RA (1992) Computerized detection of masses in digital mammograms: Automated alignment of breast images and its effect on bilateral-subtraction technique. Medical Physics pp

[3] Kupinski M, Giger ML, Lu P, Huo Z (1995) Computerized detection of mammographic lesions: Performance of artificial neural network with enhanced feature extraction. Proc SPIE 2434, pp 598-605.

[4] Chan H-P, Doi K, Vyborny CJ, Metz CE, MacMahon H, Jokich P, Galhotra S (1987) Digital mammography: Development of a computer-aided system for detection of microcalcifications. Proc SPIE 767, pp 367-370.

[5] Chan H-P, Doi K, Vyborny CJ, Lam KL, Schmidt RA (1988) Computer-aided detection of microcalcifications in mammograms: Methodology and preliminary clinical study. Investigative Radiology 23, pp 664-671.

[6] Chan H-P, Doi K, Vyborny CJ, Schmidt RA, Metz CE, Lam KL, Ogura T, Wu Y, MacMahon H (1990) Improvement in radiologists' detection of clustered microcalcifications on mammograms: The potential of computer-aided diagnosis. Investigative Radiology 25, pp 1102-1110.

[7] Ema T, Doi K, Nishikawa RM, Jiang Y, Papaioannou J (1995) Image feature analysis and computer-aided diagnosis in mammography: Reduction of false-positive clustered microcalcifications using local edge-gradient analysis. Medical Physics 22, pp 161-169.

[8] Nishikawa RM, Giger ML, Doi K, Vyborny CJ, Schmidt RA (1995) Computer-aided detection of clustered microcalcifications using digital mammograms. Medicine and Biology in Engineering and Computing 33, pp 174-178.

[9] Nishikawa RM, Jiang Y, Giger ML, Schmidt RA, Vyborny CJ, Zhang W, Papaioannou J, Bick U, Nagel RH, Doi K (1994) Performance of automated CAD schemes for the detection and diagnosis of clustered microcalcifications. In: Gale AG, Astley SM, Dance DR, Cairns AY, (eds.), Digital Mammography. Elsevier Publishing, Amsterdam, pp 13-20.

[10] Zhang W, Doi K, Giger ML, Nishikawa RM, Schmidt RA (1996) An improved shift-invariant artificial neural network for computerized detection of clustered microcalcifications in digital mammograms. Medical Physics 23, pp 595-601.

[11] Schmidt RA, Nishikawa RM, Osnis R, Schreibman KL, Giger ML, Doi K (1996) Computerized detection of lesions missed by mammography. In: Doi K, Giger ML, Nishikawa RM, Schmidt RA, (eds), Digital Mammography '96. Elsevier Science, Amsterdam, pp 105-110.

[12] te Brake GM, Karssemeijer N, Hendriks JHCL (1998) Automated detection of breast carcinomas not detected in a screening program. Radiology 207, pp 465-471.

Acknowledgments

The authors thank Augustine M Urbas, who helped in the initial data collection.

This work was supported in part by grants from the NIH (CA60187 and T32 CA09649) and the US Army (DAMD17-96-1-6058, DAMD17-96-1-6228). The contents of this paper are solely the responsibility of the authors and do not necessarily represent the official views of any of the supporting organizations. Robert M. Nishikawa, Maryellen L. Giger, Robert A. Schmidt, and Kunio Doi are shareholders in R2 Technology, Inc. (Los Altos, CA). It is the University of Chicago Conflict of Interest Policy that investigators disclose publicly actual or potential significant financial interests which may appear to be affected by the research activities.

Running title

CAD ANALYSIS OF 10,000 CASES

Taken from: Digital Mammography 2000, Yaffe MJ, editor (Medical Physics Publishing, Madison WI, 2000).

Prospective Computer Analysis of Cancers Missed on Screening Mammography

Robert M. Nishikawa, Maryellen L. Giger, Robert A. Schmidt,*

Carl J. Vyborny, Ulrich Bick, and Kunio Doi

Department of Radiology, The University of Chicago

*Department of Radiology, New York University

1. Introduction

For a computer-aided detection (CAD) scheme to be an effective aid to radiologists, two conditions must be met. First, the computerized detection scheme must be able to detect cancers that a radiologist would overlook. Second, the radiologist when using the aid must act appropriately (i.e., correctly dismiss computer false positives and call back women with cancer). While at least three studies have indicated that automated detection schemes can find cancers missed on mammograms (Schmidt *et al.*, 1996; te Brake *et al.*, 1998; Warren-Burhenne *et al.*, 2000), these were all done using cases selected retrospectively. In this paper, we expand our study of the first requirement – that the computer can detect cancers overlooked on a screening mammogram – in a prospective study.

We previously reported on our prospective study of computerized detection of cancers on screening mammograms. We found that approximately 50% of cancers missed on a screening mammogram that are apparent in retrospect can be detected by one of our automated detection schemes (Nishikawa *et al.*, 1999). Visually, some of the overlooked cancers were very subtle and did not appear very different from normal breast tissue. In this study, we determined what fraction of these cancers are detectable in a screening-type environment.

Taken from: Digital Mammography 2000, Yaffe MJ, editor (Medical Physics Publishing, Madison WI, 2000).

2. Materials and Method

Cases and computer outputs used in this study were collected from a prospective study of CAD for screening mammography. At the University of Chicago Hospitals, we have been digitizing all screening mammograms since November 10, 1994. To identify which women in our study cohort have developed breast cancer, we compared the list of all patients included in our study against all breast pathology reports from our Hospital. For all women who had breast cancer, we examined all of their screening mammograms that were included in our study, along with diagnostic exams and, in some cases, needle localization exams. In this way, we were able to identify all cases where a cancer was visible on a screening mammogram. In some cases, these screening mammograms were read as abnormal and the women were called back, and in others, the cancer was overlooked and the mammogram was called normal. Here, we refer to the latter as a missed cancer.

To determine what fraction of these missed cancers can be detected in a screening environment, we conducted an observer study. We asked three radiologists to read 75 screening cases in which the cases containing missed cancers (n=21) were mixed with exams that contained a screen-detected cancer (n=3) and cases without cancer (n=51). The cases were presented in random order on a mammography motorized viewer. Magnifying glasses were available. No time limit was imposed.

The three radiologists were all specialists in breast imaging. Two had over 15 years experience and are MQSA qualified. The third, a European radiologist, with over 10 years of experience, had extensive experience in breast imaging, including digital mammography and breast MRI.

Taken from: Digital Mammography 2000, Yaffe MJ, editor (Medical Physics Publishing, Madison WI, 2000).

For each case, we included previous exams, when they were available. For each case, we asked the radiologist to give their BI-RADS assessment. Based on this assessment, we determined what fraction of radiologists would call back the cases containing a missed cancer. We also asked the radiologists to give their level of confidence that the patient should be called back for further imaging or for a biopsy. This was done using a visual analog scale with the left end marked as "definitely do not call back" and the right end marked as "definitely call back". The observers were instructed that if they were equivocal about calling the patient back, then they should mark the center of the scale. Short-term follow-up did not count as call back.

Two different detection schemes were used in this study: one for the detection of masses and the other for the detection of clustered microcalcifications. Details of these schemes have been described previously (Bick *et al.*, 1995; Nishikawa *et al.*, 1995; Yin *et al.*, 1993; Zhang *et al.*, 1996). Our prospective study began in November, 1994. The algorithms used throughout the study were kept constant, so those 1994 versions were used. Since then, the false-positive rate has been reduced, but these newer techniques have not been incorporated into the system yet (Anastasio *et al.*, 1998; Kupinski and Giger, 1998; Yoshida *et al.*, 1996).

3. Results

In the first three years of our study, 12,670 exams, which were obtained from 9195 women, were analyzed on our CAD workstation. Of these women, 79 developed breast cancer (minimum two years of follow-up). Sixty-one of the cancers were detected on a screening mammogram. The rest were detected on a diagnostic mammogram, or were palpable or both. Sixty-five cancers were visible mammographically. In the 79 cancer cases, 42 cases had a negative screening

mammogram that was included in our study. Of the 42, 19 were mammographically occult in retrospect and 23 had a lesion that was visible at the site where the cancer developed. Examining the prospective computer results for those 23 cases showed that 12 of these cancers were detected by the computer.

All 12 of the computer-detected, radiologist-missed cancers and 9 of the 11 computer-missed, radiologist-missed cancers were used in the observer study. Two computer-missed, radiologist-missed cancer cases were not available for the study. Added to these 21 cases were 3 randomly selected screen-detected cases and 51 normal cases (based on at least two-year follow up) for a total of 75 cases. The normal cases were selected randomly from patients who had a screening mammogram in 1995 and at least one additional exam at least a two years later. Computer sensitivity on the cancer cases used in this study was 62.5% (15/24) and the false positive-rate on all 75 cases was 0.9 per image for calcifications and 2.1 per image for masses.

From the rating data, ROC curves were plotted (see Figure 1). In addition, using the BI-RADS assessment, we determined the sensitivity and specificity for each reader. These are shown as letters in Figure 1 and are reported in Table 1. Also listed in Table 1 are the sensitivity and specificity for the computer schemes and for the clinical interpretation of the screening case. The computer had at least one detection in each case and thus had a specificity of zero. The clinical readings had 100% specificity since the normal cases were found based on a normal screening mammogram. Similarly, the sensitivity of the clinical readings was low since we intentionally included exams where a cancer was overlooked. Note, however, that one of the cases detected clinically was missed by one of the three radiologists.

Taken from: Digital Mammography 2000, Yaffe MJ, editor (Medical Physics Publishing, Madison WI, 2000).

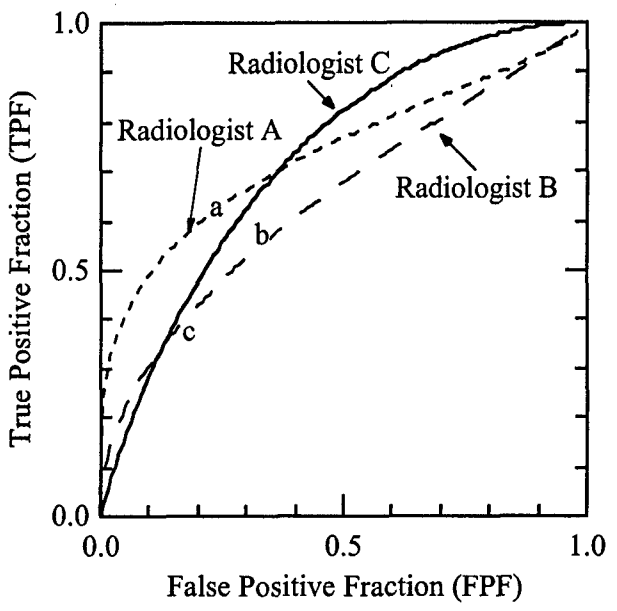

Figure 1. ROC curves for the 3 readers. The lower case letters indicate the operating points (sensitivity and specificity) as determined by their BI-RADS assessment. The areas under the ROC curves, A_z, $\pm$ one standard deviation were 0.73 ± 0.06, 0.64 ± 0.07, and 0.73 ± 0.06 for radiologists A, B, and C respectively.

Based on the BI-RADS assessment, we determined the number of times a case was given a 0, 4, or 5 score (call back or biopsy). We then compared the computer performance of sub categories of the data based on the number of times the cases were called abnormal. We included the clinical reading in this analysis, so that there were four assessments made per case (see Table 2). Of the 12 cancer cases that were missed clinically and detected by the computer, 7 were detected by 2 of the 3 readers in this study and 11 were detected by at least one of the readers. One the other hand, some of the computer-detected cancers are below the detection threshold of experienced radiologists – 4 of the 12 cancers were not detected by any of the radiologists.

Taken from: Digital Mammography 2000, Yaffe MJ, editor (Medical Physics Publishing, Madison WI, 2000).

Table 1. Sensitivity and specificity for the three readers, the clinical reading and the computer.

Reader	Sensitivity	Specificity
A	63%	76%
B	58%	67%
C	38%	82%
Clinical	13%	100%
Computer	63%	0%

4. Discussion and Conclusions

The data in Table 2 show that the computer can detect cancers that are missed by a radiologist and the majority of those computer-detected missed cancers are detectable by a radiologist. When either 2 or 3 of 4 radiologists detected the cancer, the computer had high sensitivity, 89% (8/9). This is in spite of the fact that the overall sensitivity of our two computer schemes is approximately 70% for all cancer cases in our prospective study [Nishikawa, 1999 #5].

A possible drawback of CAD is that computer could increase the call-back rate. Approximately 80% of lesions identified by a radiologist in a normal mammogram were also identified by the computer as a potential lesion. In the same way that we infer that radiologists detecting missed

Table 2. Number of radiologists recommending call back or biopsy for the normal and cancer cases. Also include is the computer performance on those cases.

# of Radiologists Recommending Call Back	Normal Cases	Cancer Cases	# of Cancer Cases Detected by Computer	Computer Sensitivity
0/4	24	4	1	25%
1/4	17	9	4	44%
2/4	9	3	3	100%
3/4	1	6	5	83%
4/4	0	2	2	100%
Total	51	24	15	62%

cancers can lead to improved sensitivity, the high correlation of false-positive lesions between radiologists and the computer would indicate that the call-back rate *may* increase with implementation of CAD. This needs to be confirmed in clinical evaluations. One initial study found no increase in call-back rate when CAD was introduced (Warren-Burhenne *et al.*, 2000). However, the study did not report on whether sensitivity increased with CAD.

Increased call-back rate with CAD must be kept in context. Currently between 5-15% of all screening exams are considered abnormal and the patient is called back for further imaging studies. Since the cancer prevalence rate in a screening population is only 0.5%, approximately 10 to 30 women are called back for every cancer detected. If CAD can detect what would have been otherwise a missed cancer for every 10-30 extra women called back because of CAD, then the "cost/benefit ratio" remains unchanged, but a cancer would have been detected at an earlier stage. Because it is difficult to differentiate benign from malignant lesions mammographically, it is not reasonable to expect CAD to increase sensitivity, without increasing the number of call backs.

The data presented in this paper provide some evidence that computer-detected cancers can help radiologists avoid overlooking cancers. We plan to conduct an observer study to determine the number of cancers initially missed by a reader that are detected when the computer results are available. To determine the actual benefits and costs of using CAD, clinical trials need to be performed. As more systems become commercially available and more widely disseminated, these questions can readily be answered.

5. Acknowledgments

The authors thank Charles Metz for use of his ROC software. This work was done as part of the International Digital Mammography Development Group (IDMDG), and was supported in part by grants from the NCI (CA 60187 and T32 CA09649, the US Army (DAMD17-96-1-6058 and DAMD17-96-1-6228). The contents of this paper are solely the responsibility of the authors and do not necessarily represent the official views of any of the supporting organizations. All authors are shareholders in R2 Technology, Inc. (Los Altos, CA). It is the

University of Chicago Conflict of Interest Policy that investigators disclose publicly actual or potential significant financial interests that may appear to be affected by the research activities.

6. References

Anastasio, M. A., H. Yoshida, R. Nagel, R. M. Nishikawa, and K. Doi. 1998. A genetic algorithm-based method for optimizing the performance of a computer-aided diagnosis scheme for detection of clustered microcalcifications in mammograms. *Medical Physics* 25:1613-1620.

Bick, U., M. L. Giger, R. A. Schmidt, R. M. Nishikawa, D. E. Wolverton, P. Lu, C. J. Vyborny, and K. Doi. 1995. Automated segmentation of digitized mammograms. *Acad. Rad.* 2:1-9.

Kupinski, M., and M. L. Giger. 1998. Automated seeded lesion segmentation on digital mammograms. *IEEE Transactions on Medical Imaging* 17:510-517.

Nishikawa, R. M., M. L. Giger, K. Doi, C. J. Vyborny, and R. A. Schmidt. 1995. Computer-aided detection of clustered microcalcifications using digital mammograms. *Medicine and Biology in Engineering and Computing* 33:174-178.

Nishikawa, R. M., M. L. Giger, R. A. Schmidt, D. E. Wolverton, and K. Doi. 1999. Prospective testing of a clinical CAD workstation for the detection of breast lesions on mammograms. In: *Computer Aided Diagnosis in Medical Imaging.* K. Doi, H. MacMahon, M. L. Giger and K. R. Hoffmann, eds. Amsterdam: Elsevier.

Schmidt, R. A., R. M. Nishikawa, R. Osnis, K. L. Schreibman, M. L. Giger, and K. Doi. 1996. Computerized detection of lesions missed by mammography. In: *Digital Mammography '96.* K. Doi, M. L. Giger, R. M. Nishikawa and R. A. Schmidt, eds. Amsterdam: Elsevier Science.

te Brake, G. M., N. Karssemeijer, and J. H. C. L. Hendriks. 1998. Automated detection of breast carcinomas not detected in a screening program. *Radiology* 207:465-471.

Warren-Burhenne, L. J., S. A. Wood, C. J. D'Orsi, S. A. Feig, D. B. Kopans, K. O'Shaughnessy, E. A. Sickles, L. Tábar, C. J. Vyborny, and R. A. Castellino. 2000. Potential contribution of computer-aided detection to the sensitivity of screening mammography. *Radiology* 215:554-562.

Yin, F. F., M. L. Giger, C. J. Vyborny, K. Doi, and R. A. Schmidt. 1993. Comparison of bilateral-subtraction and single-image processing techniques in the computerized detection of mammographic masses. *Invest Radiol* 28:473-81.

Yoshida, H., K. Doi, R. M. Nishikawa, M. L. Giger, and R. A. Schmidt. 1996. An improved computer-assisted diagnostic scheme using wavelet transform for detecting clustered microcalicifications in digital mammograms. *Academic Radiology* 3:621-627.

Zhang, W., K. Doi, M. L. Giger, R. M. Nishikawa, and R. A. Schmidt. 1996. An improved shift-invariant artificial neural network for computerized detection of clustered microcalcifications in digital mammograms. *Medical Physics* 23:595-601.

Can computer-aided diagnosis (CAD) help radiologists find mammographically missed screening cancers

Robert M. Nishikawa*[a], Maryellen L. Giger[a], Robert A. Schmidt[b], John Papaioannou[a]
[a]Department of Radiology, The University of Chicago; [b]Department of Radiology, New York University

ABSTRACT

We present data from a pilot observer study whose goal is design a study to test the hypothesis that computer-aided diagnosis (CAD) can improve radiologists' performance in reading screening mammograms. In a prospective evaluation of our computer detection schemes, we have analyzed over 12,000 clinical exams. Retrospective review of the negative screening mammograms for all cancer cases found an indication of the cancer in 23 of these negative cases. The computer found 54% of these in our prospective testing. We added to these cases normal exams to create a dataset of 75 cases. Four radiologists experienced in mammography read the cases and gave their BI-RADS assessment and their confidence that the patient should be called back for diagnostic mammography. They did so once reading the films only and a second time reading with the computer aid. Three radiologists had no change in area under the ROC curve (mean A_z of 0.73) and one improved from 0.73 to 0.78, but this difference failed to reach statistical significance (p=0.23). These data are being used to plan a larger more powerful study.

Keywords: computer-aided diagnosis, mammography, screening, missed cancers, observer study, breast cancer

1. INTRODUCTION

Computer-aided diagnosis (CAD) has the potential to improve the accuracy of mammography by reducing the number of cancers overlooked by a radiologist reading without a computer aid. For CAD to be effective, two conditions must exist. First, the computer must be capable of detecting cancers that a radiologist would overlook. Second the radiologist must be able to recognize when the computer has detected an overlooked cancer and call the patient back for work up, while at the same time determining when the computer has identifed a false positive (i.e. non-malignant lesion). The long-term goal of our research is to show that both these conditions can exist clinically.

To show that automated detection schemes can detect overlooked cancers, we have conducted a prospective study running our two detection schemes on over 25,000 consecutive screening mammograms. We showed, using data collected in that prospective study, that our detection schemes can detect approximately 50% of cancers overlooked in a screening mammogram, but were visible in retrospect.[1] This result is consistent with retrospective studies of CAD and missed cancers.[2-4] In the current study, we report on a pilot study design to collect data to plan for a large full-scale observer study that will measure radiologists' ability to effective use the computer aid.

2. MATERIALS AND METHODS

2.1 CAD Schemes

We have developed two detection schemes: one for cluster microcalcifications and the other for masses, or more precisely, any non-calcific lesion. These have been described previously.[5-15] These two schemes have undergone clinical evaluation. The version of the schemes used for that study were the ones that existed when the study began in November 1994. We used the same schemes throughout the study, even though substantial improvements have been made since then. Figures 1 and 2 show flowcharts of the schemes.

2.2 Case Selection

The cases used in this study were drawn from our clinical evaluation. Over 25,000 screening mammograms have been analyzed by the two schemes. We have follow-up data on the first 12,690 cases. A total of 79 women in that group

*r-nishikawa@uchicago.edu; Dept. of Radiology, The University of Chicago, 5841 S Maryland Ave. MC-2026, Chicago IL 60637

Image Perception and Performance, Elizabeth A. Krupinski, Dev P. Chakraborty, Editors, Proceedings of SPIE Vol. 4324 (2001)

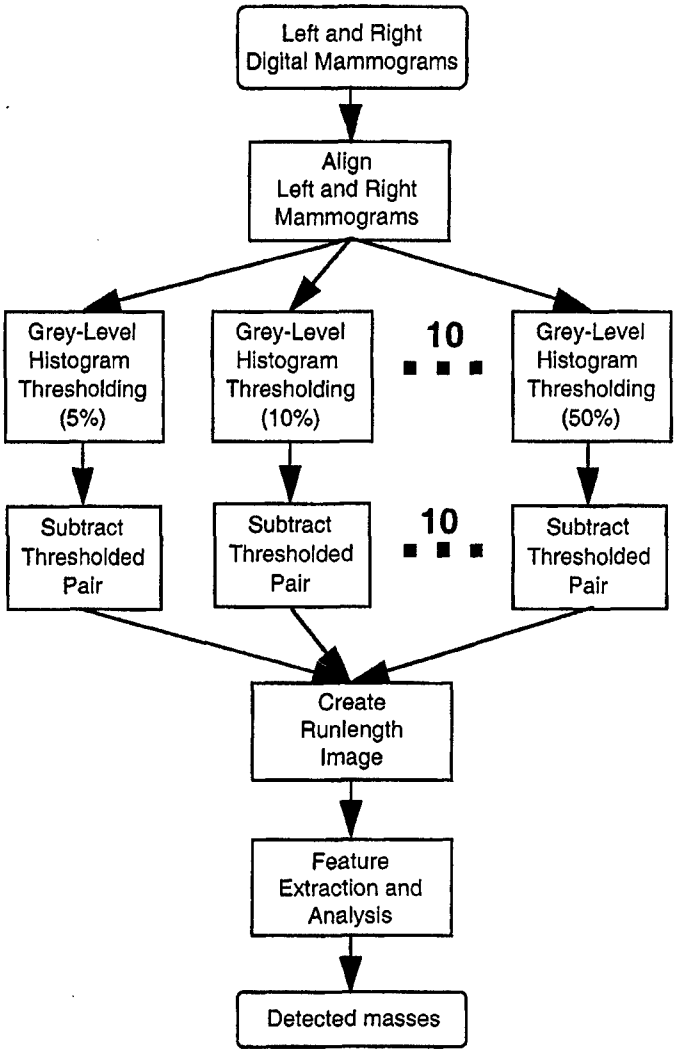

Figure 1. Flowchart of mass detection scheme.

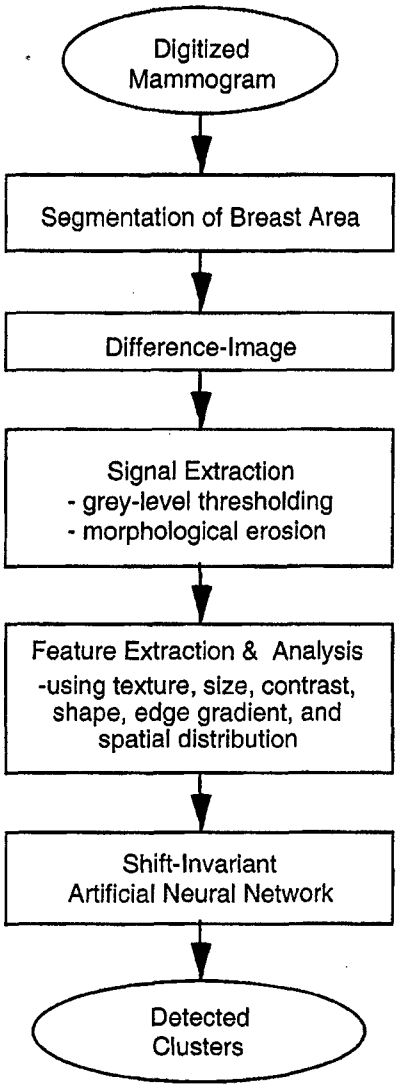

Figure 2. Flowchart of clustered microcalcification detection scheme.

developed breast cancer (after at least two years of local follow-up). Of those women, 42 had a negative screening mammogram that was included in our study cohort. We reviewed the negative screening mammograms for those women and determined that in 23 of the cases, some indication of the cancer was present on the negative mammogram. Review of the computer detections for those 23 cases showed that in 12 cases the computer identified the cancer that had been overlooked clinically.

For our observer study, we chose 75 cases from the 12,690. Figure 3 shows schematically the breakdown of cases. We included all 12 cases with a computer-detected cancer on a negative mammogram, 9 of 11 cases of a computer-missed cancer on a negative mammogram (2 cases were unavailable for the study), 3 clinically-detected cancer cases (for which the computer detected the cancer also) and 51 normal cases (cancer free after at least two years). In total, there were 24 cancers of which the computer detected 15, although only three were detected clinically. The computer false-positive rate for these 75 cases was 2.0 per image for masses and 0.96 per image for clustered calcifications.

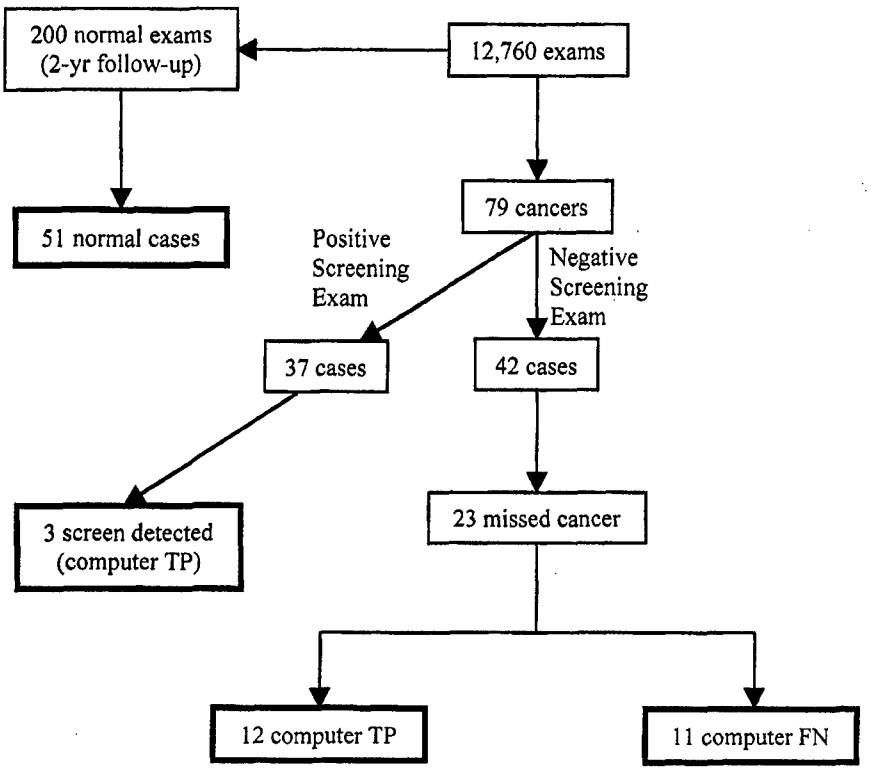

Figure 3. A breakdown of the cases used in this pilot observer study taken from a study of 12,690 clinical screening cases subjected to computer analysis.

2.3 Observers

There were four observers in this study. Three were American radiologists with over 15 years experience in breast imaging and were MQSA-qualified. The other radiologist was from Europe with over 10 years of experience in breast imaging. All readers had some experience with CAD, either in research or clinically or both.

2.3 Study Design

For each case, under each condition (aided and unaided), the readers answered two questions. The first was to give their BI-RADS assessment of the case (see Table 1). The second was to give their confidence that the patient needed work-up or a biopsy, that is, the woman had a positive screening mammogram. There confidence rating was recorded on a visual-analog scale, which required the radiologist to place a mark on a 5-cm line. The left end of the line was labeled "Definitely Do Not Call Back" and the right end of the line was labeled "Definitely Call Back."

We used a sequential reading design in which the readers first read the cases without the computer aid and gave their opinion. The readers were then immediately shown the computer's detection output, and gave another opinion. This method differs from the conventional method of performing observer studies in which the two different reading conditions – unaided and aided, in this case – are rendered on different days for a given case. We chose to use the sequential method because it more closely resembles the way in which CAD is used clinically. The conventional method is an artificial construct unrelated to the clinical use of CAD. So while in the sequential method, the aided condition is always second, which would normally create a bias, when CAD is implemented clinically, the aided opinion will be second, after the radiologist first reads the images without aid. The other advantage of this approach is that it gives more statistical power to the experiment as will be shown in the Discussion section.

Table 1. Definition of the BI-RADS assessment categories. In this study, categories 0, 4, and 5 were considered positive exams (abnormal mammograms) and 1, 2, and 3 were considered negative exams (normal mammograms). This table was taken from reference [16].

Category 0: Need Additional Imaging Evaluation. Finding for which additional imaging evaluation is needed. This is almost always used in a screening situation and should rarely be used after a full imaging work up. A recommendation for additional imaging evaluation includes the use of spot compression, magnification, special mammographic views, ultrasound, etc.
Category 1: Negative. There is nothing to comment on. The breasts are symmetrical and no masses, architectural disturbances or suspicious calcifications are present.
Category 2: Benign Finding. This is also a negative mammogram, but the interpreter may wish to describe a finding. Involuting, calcified fibroadenomas, multiple secretory calcifications, fat containing lesions such as oil cysts, lipomas, galactoceles, and mixed density hamartomas all have characteristic appearances, and may be labeled with confidence. The interpreter might wish to describe intramammary lymph nodes, implants, etc. while still concluding that there is no mammographic evidence of malignancy.
Category 3: Probably Benign Finding - Short Interval Follow-Up Suggested. A finding placed in this category should have a very high probability of being benign. It is not expected to change over the follow-up interval, but the radiologist would prefer to establish its stability. Data are becoming available that shed light on the efficacy of short interval follow-up. At the present time, most approaches are intuitive. These will likely undergo future modification as more data accrue as to the validity of an approach, the interval required, and the type of findings that should be followed.
Category 4: Suspicious Abnormality - Biopsy Should Be Considered. These are lesions that do not have the characteristic morphologies of breast cancer but have a definite probability of being malignant. The radiologist has sufficient concern to urge a biopsy. If possible, the relevant probabilities should be cited so that the patient and her physician can make the decision on the ultimate course of action.
Category 5: Highly Suggestive of Malignancy - Appropriate Action Should Be Taken. These lesions have a high probability of being cancer.

Radiologists read original mammograms mounted on a mammographic motorized viewer. In 73% of the cases, when they were available, a previous screening mammogram was also mounted on the viewer. The average time between the current and previous exams was 1.6 years for both the normal and abnormal cases. A computer interface was provided to collect the readers' responses. When the observer rate the case as abnormal, they specified the location of all suspicious lesions on a digital copy of the exam that was displayed on the computer interface. This interface was also used to display the results of the computerized detection schemes to the observers. A magnifier was used by the radiologists and there was no time limited imposed.

3. RESULTS

Using the confidence rating scale, ROC curves were generated for the four readers under the two reading conditions (without and with computer aid). These are shown in Figure 4. The areas under the curve values are given in Table 2. Only one of the four radiologists had an improvement in performance (Reader B), but the improvement was not statistically significant ($p=0.23$). Overall, with aid, 3 extra cancers were detected by the radiologists. This was out of 49 undetected cancers (summed over all four readers) in the unaided reading condition. The computer indicated the cancer in 20 of these 49 cases. For the normal cases, the computer caused 9 extra call backs by the radiologists. In the unaided condition, there were a total of 51 false-positive call backs by the four readers and 153 true-negative calls. Twenty-four of the computer false positives match one of the 51 false positives by the radiologists.

4. DISCUSSION AND CONCLUSIONS

While the computer correctly identified the cancer in 20 of 49 cancers missed by the radiologists (summed over the four readers) in the unaided reading condition, only 3 extra cancers were detected by the readers in the aided reading condition. Possible reasons for this are: the computer-detected cancer was below a particular radiologist's threshold for call back, the radiologist thought that the lesion was benign, and radiologist ignored or did not considered carefully enough the computer detection. The latter is a function of the computer performance. In our experiment, the false-positive rate was fairly high

with an average of 3.0 false positives per image or 12 per case. A large number of computer false positives requires extra time for the radiologist to review the case, since the radiologist must correlate the location in the computer-output image with the corresponding location on the film mammograms and then reassess that area. The higher the false-positive rate the lower the likelihood that the radiologist will check each location thoroughly. A lower false-positive rate may lead to a higher number of cancers being detected with computer aid.

A high false-positive rate also increases the chances of the computer generating a false-positive overcall by the reader, since a computer prompt on a non-cancer could case the radiologist to call the patient back, when this would not have occurred in the absence of the computer aid. Nearly half of the radiologists' false positives (24/51) in the unaided reading condition corresponded to a computer false positive. Since in only one case did all four radiologists have the same false positive, there is a potential for a computer prompt corresponding to a false positive by one radiologist to be overcalled by a different radiologist in the aided condition. In fact, there were three times as many extra false positives compared to extra cancers detected when the computer aid was used.

This experiment was a pilot observer study. In our full observer study, we will use CAD schemes with lower false-positive rates. This should help make the computer aid more beneficial to the observers.

Table 2. Area under the ROC curves for the four readers under the two different reading conditions: unaided and aided (CAD).

Reader	Unaided	CAD
A	0.69	0.68
B	0.73	0.77
C	0.80	0.79
D	0.71	0.69

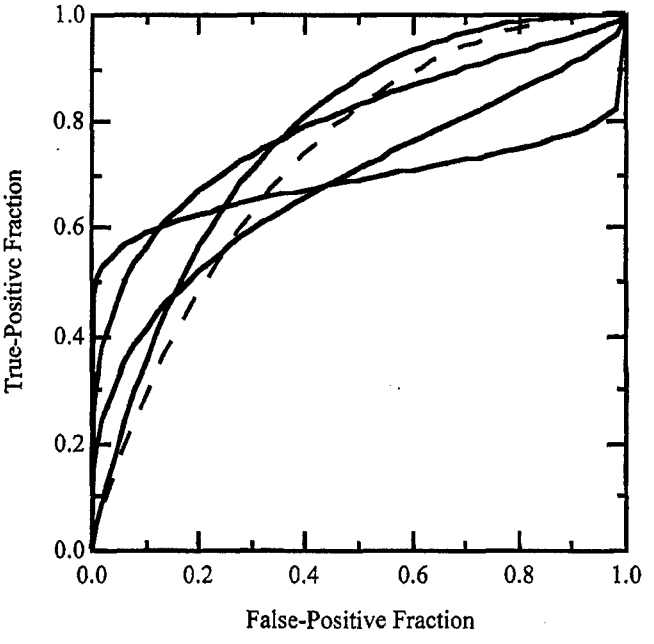

Figure 4. ROC curves for the four readers. Shown are curves for the aided reading condition (solid lines) and, for Reader B, the curve for the unaided reading conidition (dased lined). The solid curve that most closely follows the dashed curve is the aided ROC curve for Reader B. The unaided curves for the over three readers were very slimilar to their aided curves.

Another interesting observation from our pilot study was that, in the unaided reading condition, the radiologists' had a call back rate between 20%-33%, based on their BI-RADS assessment. Clinically, the call back rate is between 5%-15%. Two possible explanations for this higher call-back rate are: (1) This was a test situation and readers read more aggressively; and (2) The prevalence rate in the study was 33% compare to 0.5%. This higher prevalence resulted in the readers reading more aggressively. It is not possible from this study to determine which if either was the cause of the high call-back rate. However, in our full study, we plan to use a lower prevalence, around 20%, and stress to the observers that the penalty for missing a cancer is the same as over calling. It is important to have the observers read as they would read clinically, if we want to be able to generalize the results of our study to general clinical practice.

Finally, to simulate the way CAD would be used clinically, we used a sequential reading method. Traditionally, observer studies are performed using two independent readings, one for each reading condition. Kobayashi *et al.* performed an observer study using both methods and found no statistically significant difference between the two reading methods, based on area under the ROC curve (A_z).[17] The biggest advantage of the sequential method is that for the same number of readers and cases, the experiment has more statistical power. Because the reader gives scores for both reading condition in the same reading sitting, those values are correlate and the A_z values are highly correlated – ρ is the correlation between A_z values in the aided and unaided reading condition. Figure 5 shows the dependence of power on ρ. In our experiment, the reader who had the largest improvement in A_z had a ρ of 0.82. The other three readers had an average ρ of 0.94. These correlation values are high because the readers had complete knowledge of their unaided score when they gave their aided score. In the independent reading method ρ is more typically 0.4, because of intra-observer variability.[18] Figure 6 compares power curves for the two different values of ρ. Even if the prevalence is increased up to 50%, more cases would be required to get the same power as when ρ is high.

In summary, from this pilot study, we have planned a 400-case reader study to measure the ability of radiologist to reduce the number of overlooked screening cancers using CAD. Compared to our pilot study, we will reduce the cancer prevalence to 20% and use detection schemes with lower false-positive rates to improve the overall design of the study.

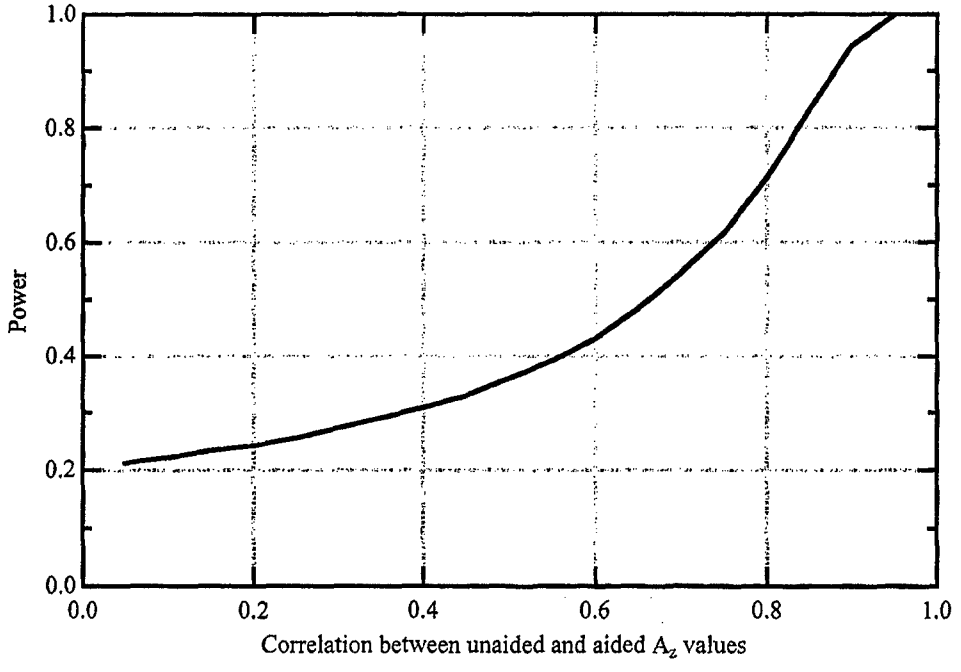

Figure 5. The dependence of statistical power on the correlation between area under the ROC values for the unaided and aided reading conditions.

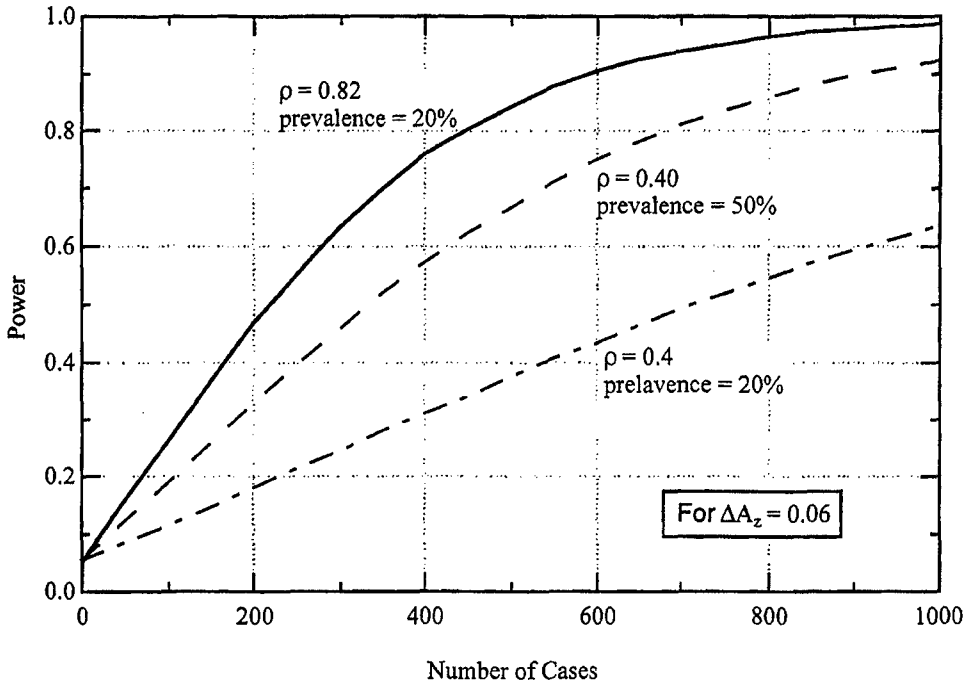

Figure 6. Power calculations for the full observer study, based on data collected in this pilot study. Curves are shown for three different levels of cancer prevalence in the dataset and for two difference levels of ρ, the correlation in the area under the ROC (A_z)curves between the unaided and aided reading conditions. Here we assume that the unaided A_z is 0.70 and the aided A_z is 0.76. These curves are for a single observer.

ACKNOWLEDGMENTS

The authors acknowledge the support of the US Army (DAMD17-96-6229 and DAMD17-96-1-6058) and the NIH/NCI (CA 60187). The contents of this paper are solely the responsibility of the authors and do not necessarily represent the official views of any of the supporting organizations.

The authors thank Dr. Charles Metz for the use of his ROCKIT suite of software tools for ROC analysis (http://xray.bsd.uchicago.edu/krl/toppage11.htm).

The authors are shareholders in R2 Technology, Inc. (Los Altos, CA). It is the University of Chicago Conflict of Interest Policy that investigators disclose publicly actual or potential significant financial interests which may appear to be affected by the research activities.

REFERENCES

1. R. M. Nishikawa, M. L. Giger, R. A. Schmidt, C. J. Vyborny, U. Bick and K. Doi, "Prospective computer analysis of cancers missed on screening clinical," in *Digital Mammography 2000,* edited by M. J. Yaffe (Medical Physics Publishing, Madison WI, 2001), in press.
2. R. A. Schmidt, R. M. Nishikawa, R. Osnis, K. L. Schreibman, M. L. Giger and K. Doi, "Computerized detection of lesions missed by mammography," in *Digital Mammography '96,* edited by K. Doi, M. L. Giger, R. M. Nishikawa and R. A. Schmidt (Elsevier Science, Amsterdam, 1996), pp. 105-110.
3. G. M. te Brake, N. Karssemeijer and J. H. C. L. Hendriks, "Automated detection of breast carcinomas not detected in a screening program," *Radiology* **207**, pp. 465-471, 1998.
4. L. J. Warren-Burhenne, S. A. Wood, C. J. D'Orsi, S. A. Feig, D. B. Kopans, K. O'Shaughnessy, E. A. Sickles, L. Tábar, C. J. Vyborny and R. A. Castellino, "Potential contribution of computer-aided detection to the sensitivity of screening mammography," *Radiology* **215**, pp. 554-562, 2000.

5. F.-F. Yin, M. L. Giger, K. Doi, C. E. Metz, C. J. Vyborny and R. A. Schmidt, "Computerized detection of masses in digital mammograms: Analysis of bilateral subtraction images," *Medical Physics* **18**, pp. 955-963, 1991.
6. F.-F. Yin, M. L. Giger, C. J. Vyborny, K. Doi and R. A. Schmidt, "Comparison of bilateral-subtraction and single-image processing techniques in the computerized detection of mammographic masses," *Investigative Radiology* **28**, pp. 473-481, 1993.
7. M. Kupinski, M. L. Giger, P. Lu and Z. Huo, "Computerized detection of mammographic lesions: Performance of artificial neural network with enhanced feature extraction," *Proc. SPIE* **2434**, pp. 598-605, 1995.
8. M. A. Kupinski and M. L. Giger, "Automated seeded lesion segmentation on digital mammograms," *IEEE Trans Med Imaging* **17**, pp. 510-517., 1998.
9. H.-P. Chan, K. Doi, S. Galhotra, C. J. Vyborny, H. MacMahon and P. M. Jokich, "Image feature analysis and computer-aided diagnosis in digital radiography. 1. Automated detection of microcalcifications in mammography," *Medical Physics* **14**, pp. 538-548, 1987.
10. H.-P. Chan, K. Doi, C. J. Vyborny, K. L. Lam and R. A. Schmidt, "Computer-aided detection of microcalcifications in mammograms: Methodology and preliminary clinical study," *Investigative Radiology* **23**, pp. 664-671, 1988.
11. H.-P. Chan, K. Doi, C. J. Vyborny, R. A. Schmidt, C. E. Metz, K. L. Lam, T. Ogura, Y. Wu and H. MacMahon, "Improvement in radiologists' detection of clustered microcalcifications on mammograms: The potential of computer-aided diagnosis," *Investigative Radiology* **25**, pp. 1102-1110, 1990.
12. T. Ema, K. Doi, R. M. Nishikawa, Y. Jiang and J. Papaioannou, "Image feature analysis and computer-aided diagnosis in mammography: Reduction of false-positive clustered microcalcifications using local edge-gradient analysis," *Medical Physics* **22**, pp. 161-169, 1995.
13. R. M. Nishikawa, M. L. Giger, K. Doi, C. J. Vyborny and R. A. Schmidt, "Computer-aided detection of clustered microcalcifications: An improved method for grouping detected signals," *Medical Physics* **20**, pp. 1660-1666, 1993.
14. R. M. Nishikawa, M. L. Giger, K. Doi, C. J. Vyborny and R. A. Schmidt, "Computer-aided detection of clustered microcalcifications using digital mammograms," *Medicine and Biology in Engineering and Computing* **33** pp. 174-178, 1995.
15. W. Zhang, K. Doi, M. L. Giger, R. M. Nishikawa and R. A. Schmidt, "An improved shift-invariant artificial neural network for computerized detection of clustered microcalcifications in digital mammograms," *Medical Physics* **23**, pp. 595-601, 1996.
16. American College of Radiology, *Breast Imaging Reporting and Data System.* American College of Radiology, Reston, Virginia, 1995.
17. T. Kobayashi, X.-W. Xu, H. MacMahon, C. E. Metz and K. Doi, "Effect of a computer-aided diagnosis scheme on radiologists' performance in detection of lung nodules on radiographs.," *Radiology* **199**, pp. 843-848, 1996.
18. C. A. Beam, D. C. Sullivan and P. M. Layde, "Effect of human variability on independent double reading in screening mammography," *Academic Radiology* **3**, pp. 891-897, 1996.